DOCUMENT RESUME

ED 251 513 TM 850 020

AUTHOR Herman, Joan; And Others
TITLE 1984 Policy Studies: The Use of Testing and Evaluation for Assessing Educational Quality and Improving School Practice. Research into Practice Project.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Nov 84
GRANT NIE-G-84-0112-P4
NOTE 76p.; For summaries of these papers, see TM 850 008. Papers presented at a conference sponsored jointly by the UCLA Center for the Study of Evaluation and the UCLA Laboratory in School and Community Education (Santa Monica, CA, June 7, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Cognitive Processes; *Educational Improvement; Elementary Secondary Education; Evaluation Methods; Evaluation Needs; *Evaluation Utilization; Feedback; Inservice Teacher Education; Multiple Choice Tests; Outcomes of Education; *Test Use

ABSTRACT

The papers in this monograph address an issue of importance to educational policy and practice: the use of testing and evaluation to assess the quality of education and to facilitate school improvement. The authors consider the traditional role that testing has played in accountability and the role that assessment and evaluation can and should play in improving teaching and learning; they point out some of the problems and limits of current evaluation practices and call for new approaches that will broaden perspectives on schooling and contribute to the usefulness of the evaluation enterprise. The four papers are: (1) "Evaluating Educational Quality: A Rational Design," by Eva L. Baker; (2) "Beyond Outcome Measures: An Agenda for School Improvement," by John Goodlad; (3) "Using Educational Evaluation for the Improvement of California Schools," by Elliot Eisner; and (4) "The Influence of Testing on Teaching and Learning," by Norman Frederiksen. (BW)

DELIVERABLE - NOVEMBER 1984

RESEARCH INTO PRACTICE PROJECT

1984 Policy Studies: The Use of Testing and Evaluation for Assessing Educational Quality and Improving School Practice

Joan Herman

Project Director

Grant Number

NIE-G-84-0112, P 4

Center for the Study of Evaluation

Graduate School of Education

University of California, Los Angeles

The project presented or reported herein was supported in part pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

INTRODUCTION

The papers in this monograph address an issue of importance to educational policy and practice: the use of testing and evaluation to assess the quality of education and to facilitate school improvement. Authors consider the traditional role that testing has played in accountability and the role that assessment and evaluation can and should play in improving teaching and learning; they point out some of the problems and limits of current evaluation practices and call for new approaches that will broaden perspectives on schooling and contribute to the usefulness of the evaluation enterprise.

The papers are drawn from "Wagging the Dog, Carting the Horse: Testing vs. Improving California's Schools," a conference sponsored jointly by the UCLA Center for the Study of Evaluation and the UCLA Laboratory in School and Community Education, both units within the Graduate School of Education. The conference was held on June 7, 1984 and attracted a diverse audience of over 100 educational practitioners, policy-makers and researchers. The conference contributed substantially to promoting dialogue and communication among these various groups, interactions which help to bridge the gap between research and practice.

EVALUATING EDUCATIONAL QUALITY: A RATIONAL DESIGN

Eva L. Baker

UCLA Center for the Study of Evaluation

The Promise

The world is too capricious for us to accept it as it is. So for psychological as well as practical reasons we have come to believe that we can influence the course of events. Large numbers of people during great epochs of history did not so believe. Whatever occurred was accepted as predestined either because of an unknowable master plan, or as a consequence of behavior in former incarnations.

Times have changed. Reasoning and thought have come to have specific uses in the information-driven society of the present. We want to be rational so we can believe that we understand and control events. We plan. We implement. We assess. Then we try to learn from experience and plan better next time. The evaluation process, in schools and elsewhere, is based upon this view of the world. A corollary to this perspective is our focus on goals and standards. If we have a clear idea of what we want, communicate it well to all actors, and have a criterion for judgment, then we should not only see change, but the change should be in the direction intended. Obvious stuff for school people who have had a surfeit of experience with goals, objectives, and standards. It should be easy; it should work. It isn't, and it doesn't very often. The purpose of this

paper is to describe what goes wrong with the good idea of evaluation for school improvement and to suggest some possible remedies.

The Problem

Schools have had years of experience with evaluation. Sometimes the function has been called testing, grading, standards or assessments and has been applied to student performance. Later, these activities were directed at programs as well as at people. So school people are not newcomers to the evaluation, although they may neither know nor particularly care about the newest name applied. They have experienced waves of equity, quality, improvement, in criterion-referenced, norm-referenced, goal-oriented, decision-theoretic, responsive, goal-free, illuminative, descrepant, creative, and transactional methods and are now caught up in a wave of excellence. In fact, academics have perpetuated untold numbers of evaluation models and measurement approaches, activity appropriate for our personal incentive structure. Unfortunately, the one model we want remains elusive: effective. Why doesn't evaluation work the way we think it should? One might say that our expectations are too high or that the technology is weak, without much impact. The spector of bumbling educators, clinging precariously to the lowest range of the SAT (as national magazines report our performance), suggests another explanation: maybe we haven't been smart enough to figure it out. Makers of policy have acted on that belief and attempted to place the evaluation of learning in the hands of presumed betters: technicians who use either cost-benefit formulae or psychometrically elegant

models, and sometimes both. Incidentally, most of them have struck out as well.

Backing up and taking a simple view (how appropriate!) we might redefine the problem. On the one hand, everyone knows and even social science research supports the idea that information can be used to make improvements in programs of various sorts. A number of conditions must be met, however. First, the information has to be in a usable form, so that long trains of hypotheses and inferences may be avoided. Second, the information should be available to those individuals who are responsible for implementing changes. Obviously, the information should be timed so that changes can be made as needed.

Third, information should be valid. It should provide an accurate picture of the matters of interest. (Validity does not necessarily imply precision, a topic to which we will return later.) In addition, those charged with using the information must find it credible. Credibility gets built in many ways, through logic, or through association with authority mechanisms or persons (like experts). A strong way to build credibility is to allow the end users to design, create or amend the character of the information base, so that, in the metaphor of our economic system, they buy in, feel ownership, or invest in the entire process. Although these points provide only a quick picture of requirements for informations utilization, those of us involved in educational evaluation can identify immediately their implications: evaluation should be

designed, implemented, once more, and used at the principal unit of change--the school. Without itemizing reasons why schools are the appropriate unit of change, let us accept that much good research and analysis have led to this perception.

This analysis now aside, all of us know that evaluation activities as they operate in most school districts are driven by a different reality. Evaluation is a process mandated from above and often from outside of the operational management of the educational enterprise. School boards, legislatures, and state education leaders have legitimate questions about the effectiveness of education. These questions involve management, staffing, quality of services, as well as those concerned with the more traditional outputs of education, such as whether students learn, what they learn, and the larger question of how well they are prepared to function in the world.

A response to these two legitimate viewpoints implies at once that 1) evaluation should generate information useful at the point of change; and 2) evaluation should contribute to responsible oversight of the educational system. Thus, we find in these two views that premises, assumptions, present practices, and implications of evaluation seemingly conflict irreconcilably.

Point of change evaluation emphasizes the specialness of each site, that is, the unique character of each school, comprised of the particular staff, setting, students, and social context. Point of change evaluation implies recognition and attention to the particular personality of a given school. The evaluation effort needs to be

sensitive to the teachers involved, their experience, content and pedagogical expertise, views of their role, their stance toward their students and toward management. Clearly, point of change evaluation should have good information about students, information which extends beyond gross estimates of performance on commercial achievement tests or socio-economic status assignments.

Among the strongest demands is that the evaluation information be directed to matters of importance and to those susceptible to change. The particular content, goals, and learning problems facing the school should be reflected in the data collection strategies and in how progress is judged. To expand for a moment on this particular point, one would expect that the way any important educational goal is treated would be influenced greatly by its practicality in a specific environment. For instance, many schools have identified comprehension in reading as a principal goal to focus effort. Yet, what aspects of comprehension are appropriate for a given school population, or even groups of children within the school differ dramatically. Comprehension for children at one school may mean basic parsing of meaning to understand the literal content of a sentence, whereas comprehension for other children might involve relatively sophisticated inferencing. Both sets of staffs at the school sites may be working to capacity to improve reading comprehension. However, an evaluation or testing procedure that looked at absolute levels of performance would credit one school greatly over the other. Point of change evaluation would need to provide

information peculiar to each site so that the appropriate instructional consequences could be identified and applied. We will return to some of the methodological and research issues inherent in this point of view later.

On the other hand, a system useful for accountability and oversight demands almost a wholly different set of features. First, the database must have comparability so that contrasts among schools can be made. Second, the academic content areas of interest must be those that either have high priority for the public or those for which policy decisions are required. Such requirements implicitly restrict the number of measures (or indicators or constructs) employed because political priorities and policy options are definitionally constrained. A third feature of top-down assessment is the more self-conscious emphasis on the connections among organizational units and subsystems, e.g., budget, staffing, management, instruction. To summarize graphically top-down (accountability) and bottom-up (point of change) evaluation features, consult Figure 1. This chart is presented to identify salient contrasts and overlaps between a top-down and bottom-up evaluation perspectives. A brief review of this Figure 1 illustrates that demands of top-down and bottom-up evaluation overlap but also differ enormously. Such feature differences are also represented in reality by the deployment of multiple-data collection or evaluation projects.

CONTRASTS BETWEEN TOP-DOWN & BOTTOM-UP EVALUATION FEATURES

Figure 1

FEATURES	TOP-DOWN		BOTTOM-UP	
	High	Low	High	Low
Student Performance				
° Comparability between units	X			X
° Variation in response modes		X	X	
° Continuous appraisal		X	X	
° Responsiveness to setting		X	X	
° Individual differences		X	X	
° Multiple data sources		X	X	
° Turn-around demands				
° Reliability	X			X
Demographics				
° SES	X		X	
° Language(s)	X		X	
° Transiency	X		X	
Process				
° Instructional options		X	X	
° Placement		X	X	
° Homework, etc.		X	X	
° Special needs	X		X	
Additional Outcomes				
° Tolerance (measure)		X	X	
° Drug referrals	X		X	
° Self-concept	X		X	
° Absences		X	X	
° Vandalism	X		X	
° Referrals	X		X	
Subsystem				
Staffing				
° Assessment	X			X
° Certification	X			X
° In-service	X		X	
Resource Allocation				
° Instruction	X		X	
° People	X		X	
° Money	X			X

These evaluation efforts occur in essentially disjunct ways. For example, a typical school district in California might have evaluation activities relative to a range of separate requirements: 1) Superordinate demands; 2) District requirements (regular); 3) District requirements (special); 4) School imposed; 5) Classroom driven.

Figure 2

Types of Evaluation Demands

1. Subordinate demands
2. District regular
3. District special
4. School imposed
5. Classroom driven

At present, there are pitifully poor numbers of instances where any integration at all occurs among these different purposes. To expand, 1) superordinate demands are triggered and include exogenously requirements for state assessment programs, participation in National Assessment, research projects, advance placement and other scholastic tests. 2) Regular district requirements may include administration of one or more achievement test batteries implementation of tests for student certification, either high school exit examinations, grade-to-grade promotion examinations, or placement tests for identification purposes, e.g., special education, language deficits.

3) Special district evaluation efforts encompass those required for reporting to State and Federal agencies for special funding, any program specific assessment related to curriculum comparisons, the institution of new programs, and so on. 4) School imposed requirements may be those identified by the school as a particular planning goal, for instance, to improve written composition across the curriculum areas. 5) Classroom driven evaluation may include those commonplaces required for a teacher to perform according to expectations, e.g., moving students around, assigning grades, having conferences, as well as those pertinent to meta-instructional demands, e.g., checking to see if using a new set of workbooks was worth using again, self-assessing the quality of teaching, or trying to figure out a new way to deal with a common arithmetic problem the students have. Uses of information at a classroom level must necessarily be specifically relevant to the options perceived as available by the teacher to move on, and within his/her capacity to achieve. Timing may either be on an instantaneous fuse, "I need to reassign these students Thursday" or may for meta-instructional analysis and be deferred until the next time the unit is taught or shared with a colleague whose schedule is two or three weeks slower. Notice that for these five different types of information-driven applications we have focused almost entirely on student performance as the principal data source. It should be clear, however, that teachers' use of information from relatively formal tests, even those which they design themselves, is always augmented, elaborated, interpreted and modified by the wider sense they have about what students can

actually do. In the CSE study of test use (Herman, 1983), a ringing finding was that teachers don't pay much attention to traditional outcome measures as main information sources for instructional decision making. Why not? Teachers don't do so for a number of reasons. First, they ought to (but actually may not) be skeptical about the tests' validity, that is, how close the tests come to measuring what teachers think they are teaching children. Secondly, the well known problem of timing is critical. Third, teachers have informal ways of assessing comprehensively student performance, judging in-class behavior, homework, task-orientation, or student efforts on work other than standardized tests, and can draw upon the accumulated pattern of information that they develop about a student, and take into account ideally students' rhythms in progress rather than from a one-time, cross-section sample of performance.

Nonetheless, the actual practical database that teachers use can be regarded as either suspect (by pessimists) or open to improvement (by the rest of us). The matter simply rests upon how credible the teacher is to be the single source through which a wide set of data, implicit criteria, and totally unreviewed decisions get filtered. Were it were that we felt somewhat more certain about all teachers' competence to do this complicated job. But our teacher training programs have not taught them how, nor are they given many models or incentives to take this part of their task seriously, and it is, after all, difficult. Thus, the proposal for a top-down, bottom-up system is designed to be an aid to teachers as well as a more formal mechanism to assess and subsequently to improve educational quality.

What I have tried to outline is a complicated system that has nominally complicated information demands. The reality is such, however, that in most instances decisions are made in the absence of formal information and that the information getting, displaying, and bemoaning are more ceremonial acts than instrumental tactics. But let us return to the ideology of rationality discussed at the outset of my remarks. Information should help one get better at "x" enterprise. Schooling generates natural categories of information that ought to feed into a decision process. How does it work now?

To slightly exaggerate for dramatic effect, it doesn't. For every purpose identified earlier, superordinate, regular and special district requirements, and so on, separate and often many separate data collection activities (or evaluations or assessments) are conducted. Each of these costs money, adds one more ounce of general debilitation to the system, and hardly ever becomes integrated with the normal demands of running, improving, and satisfying the multiple clients of the schools. Since the information is rarely used, other than to rationalize a politically inspired decision, (or for real estate investors to use when marketing homes near "good" schools,) the cost we are incurring is intolerable. Now, as an appositive, we can develop some cost figures on a per student basis for testing and evaluation, and on an absolute basis, these costs are not high. What is worrisome is that these costs take a substantial part of the marginal funds our discretion, funds that might go for buying less outdated books, or adding a teacher here or there. Thus, spending money

for superficial activities required by the political arena annoys the Calvinist ancestors of my friends, if not my own. Puritanical yearnings aside, the political requirements for assessment, evaluation, and other indicators of good management will continue. The trick is to make them useful.

Thus, as my computer acquaintances are fond of saying, we have a top-down, bottom-up problem. Accountability looks top-down and drives the system based on needs for overall views of system operation, logically, if sometimes not practically related to policy making. Bottom-up needs, the classroom in particular, imply information access and use, but different kinds, at different pointsfor very different purposes. As of now, everyone gives tests and is involved in the "evaluation process", but it often mere is role playing. We want to make it real; to make the money spent show up in high quality educational services and in student performance that we can be proud of. The problemspace (more computerese) of attention is the juncture or the intersect between top-downess and bottom-upness. Do we start like the tunnel building children in the sandbox, burrowing first on one side then another, all the while hoping that their two hands meet in the unseen darkness? Perhaps a memorable analogy, but very bad evaluation planning. How do we align, partition, adapt, and adjust information needs and uses so that we produce the following?

1. A real information system, rather than the flotsam and jetsom we call evaluation now.

2. A system that is efficient.

3. A system that manages the reconcilliation of policy needs while maintaining the personality, integrity, and idiosyncracy of given schools.

4. An information base that will actually inform instruction.

Methods and Methodologies

We will start with the unit as the school. First, because of the good research alluded to earlier that supports the school as the unit for change, and felicitously, school districts often make policy at school levels. Next we have to decide what goes in such a system, and those decisions should be reached based upon what plausible uses there are now for information. Clearly, there is every justification for decisions for oversight, for public accountability. And surely, we want the particularistic time, person, and place bound problems to be addressed. Since we're creating something new, let's keep our options flexible while at the same time pursuing a design solution. Let's agree on basic parameters of the effort and then look from one side (top-down) and then the other (bottom-up) to see how we're coming.

Figure 3

Givens for a Functional School Based Evaluation System

1. School level a principal focus.
2. Student performance essential.
3. Comparability on some elements essential.
4. Utilization at all levels.

Let's explore with the idea of a comprehensive with an expansion of the following features: some of its information allows for cross-student and school comparisons. And, obviously there is a technical basis for comparability of such data. Some elements of the system are demanded. There are no choices and those indicators are identified by policy actors or, to respond to particular data needs, by suprordinate requirements. Let's also posit that the system has elements in it that allow for local option, quick turn-around, outcomes measured across time, multiple data sources on certain critical dimensions. Let's also include a place for quality of school life and quality of effort indicators, some measures of instructional resources and efforts and measures of process/outcome (depending upon perspective) including affective measures, indicators of parent/community support, measures of collaboration and integration among members of the school community. Also desirable measures of societal changes hitting the school, school specific indicators on vandalism, absenteeism, transiency, changes in demography of student or teacher groups, ses, etc.

Now what makes sense as a task design technique for such a comprehensive system?

1. Required features must be identified, ideally useful at all levels.

2. Slots for options need to be identified with either sets of optional indicators for any one slot, or open choices.

3. Not all slots should or could be filled during any one school year.

4. Slots should be filled so that longitudinal patterns might be discerned (multiyear to catch longer term effects).

5. Information overload needs to be avoided at all levels.

6. Data collection and entry should be easy and not time consuming.

7. Principal users (main users) should be participants in design of system generation.

8. Procedures for sampling, decomposition and aggregation should be included so that least amount of data necessary.

9. Let's not do the most sophisticated system we can; let's do the least that will work.

Operational fairy tale version 1.

Imagine a high school where the following essential data sets are required: 1) Competency Test Based Scores entry scores in reading and math. 2) Competency test scores on a district wide measures of reading and math. Blake high school also keeps track of the number of students taking advanced placement examinations in various fields, SAT scores of 12th graders and post secondary plans of seniors. Blake high school teachers think that there is a problem developing because absence rates seem to be higher. The school decides it wants to work on this area specifically. In addition, the school is concerned that it is not challenging its top students and wants to improve. Last, the school is taking the Carnegie Report seriously (Boyer, 1983) and wants to assure that its students are competent writers.

How would our fantasy system work in that environment? Let's just focus on two of the assorted problems. Absence rates need

attention. The system collects and sorts not only how many absences occur, the rate, but also the distributions, what kinds of students are absent, over what broad distribution of time, and over what particular days. Proximity of school events (football, dances), drug referrals, transiency and school demographics are plotted. It is not problem to sunmmarize these records for the district to consider as models for analysis. Obviously, patterns are reviewed, and if, hypothetically, a clear pattern develops, for instance, that absences are distributed unfairly to new students who avoid school activity days, something can be done.

Second problem, please recall, is improving writing. Assume the English Department of Blake high school manages to convince the rest of the faculty that writing is something that needed work. Minimally entered into the system could be the number of writing assignments received for any given student across classes, i.e., in science, with appropriate description (average length, type of assignment). In addition, imagine that the English teachers have heroically taught a common scoring system (analytic, of course,) to the teachers of other subjects. Thus, data for kids, entered on the micro by "computer" kids includes student code number (for privacy,) any scores on task competence, organization, concreteness, orthographic conventions, systax, usage, etc. In addition, lists of topics, resources, and assignments could be kept. At Blake High, "slots" in use involve across-time traking of absences, with some global SES correlates, as well as across time, course, task, skill

reporting of writing performance. Based on the baseline, and full entries on a 30% sample, teachers can see that students are having difficulty with task directions, that is, knowing what they should write about, rather than simply problems of expression. Teachers decide some explicit prewriting activities ought to be tried.

What minimal design modules should such a system have?

1. A KIDFILE, including identification of the student, pertinent demographics, existing essential comparable scores on standard tests. The Kidfiles should be aggreeable by grade, SES, absence rate, performance indicators, academic grades, course of study, years in the district, etc.

2. A DATA ENTRY SYSTEM, probable student-user dependent.

3. A MICRO with a hard disk.

4. Some PERSON, probably a teacher given one period release, to be in charge and to take the lead in interpretation. It is best if this person is not a math teacher; maybe a union leader, someone with good personal skills, and an excellent teacher.

5. A MECHANISM for decisions to be made on what aspect of instruction or program people want to move on, and for which they have plausible options. People may choose to focus where they suspect problems. Both mechanisms need to be tied explicitly to data with some identified milestones (time to look, sort, and interpret.)

6. A MECHANISM to delete things out of slots and switch effort to other areas.

7. A METHOD FOR REPORTING good works, either good effects, or interesting processes up the line to get credit from district.

Obviously, for this system to work, the larger organizational units will need to be responsive and supportive. A school district might have to explore how it can reduce the information burden it places on individual schools during the period of early implimentation of the system.

The district must:

- Provide incentives, rewards, credit, for such activities.
- Minimize its redundancies to use information for personnel decisions (move a principal based on data he/she generated).
- Protect privacy of school, staff and kids.
- Try not to add to essential list, without deleting something else. Provide a period of safety and protection for system trial (pilot) and distribution.
- Monitor and support.

What are the research issues inherent in such a system? Clearly, there is enough work to supply any individual's entirely scholarly career. Let's take the idea of measurement constructs and comparability as research issues and explore them in terms of the writing at Blake high school. From research, we know that writing performance varies enormously with the task selected and the particular topic about which the student has to write. Task differences include the different purposes of writing, often categorized by types of discourse, such as persuasion, exposition, narration, and so on, although even these categories have blurred boundaries. Task also varies in terms of the audience to whom the writing performance reflects general language facility, command of orthographic conventions,like punctuation and spelling, range and fluency of syntactic options, and individual differences in intelligence, experience, and other trait-like variables. Given that the orientation to a school level system implies that between school differences are large, and differences among children are also large, how could one develop a

writing assessment that is fundamentally valid for the experience, setting and instruction of individual children, and at the same time can provide a fair and comparale measure for groups of schools? Do we need to provide opportunity to write on the same topic across time periods, for longitudinal information? Well, of course, but what about practice effects? Do we need to use the same topics across grade levels (to look at growth)? Do we invoke the same scoring standards for students at different grade levels, even if they share the same task and topic assignments? How do we report cross-school comparisons when students at different schools can handle vastly different levels of task? How can we go about reporting on writing progress overall, without resorting to a general measure that is appropriate to no group assessed? Clearly, the top-down, bottom up system is not an off-the-shelf item. It is, however, one technology that, with its underlying theoretical and practical research issues, that may be worth our time. The goal may not to build this system, but to use the design problem as a way to shed new, and perhaps creative light on a dark space.

Beyond Outcome Measures: An Agenda for School Improvement

John Goodlad

Let me begin by talking a little bit about the circumstances in which we now find ourselves in the current furor over the reform of schooling in the United States. I think it does have to be placed in some perspective if we are going to respond appropriately. A good many analysts have pointed out that the decline in competence in schooling, as well as the increase in disaffection in schooling that occurred during the decade of the 70's is very closely linked with declining faith generally in our institutions and with the decline of the economy that began during that same period.

I don't think it's any surprise that the release of the report, A Nation at Risk, last year, had a comparable effect to the launching of Sputnik in 1957. We had been building up for it. If the Nation at Risk report had not focused our attention on schooling there would have been some other catalyst. The response was very similar to the response following Sputnik: that is, an immediate outcry regarding the quality of our schools, "the rising tide of mediocrity in our schools. If some other nation had imposed the condition of our schools on us it would have been comparable to an act of war." The report goes into a series of very specific recommendations regarding a longer school day, more math, more science, more technology, more discipline, better teachers, and a certain amount of pie in the sky, along with a lot of other rather quick remedies.

Very soon, there was the usual galvanic connecting of achievement test scores with school health. That is, there is a rising tide of mediocrity

in the schools and the presumed indicator, in large measure, is declining achievement test scores. Therefore, the indicator of improving school health will be a corresponding increase in achievement test scores.

I would like to submit that achievement test scores constitute a poor thermometer for judging the health of schools, just as the thermometer we use with human beings is a poor one for judging the health of human beings. Notice the response when a person's temperature rises and we get a reading showing 103 or 104 or 105 -- there is the immediate use of an antibiotic. Yet in the most serious illnesses, the closing up of the arteries or the beginning of a cancerous condition, the thermometer would tell us nothing. And you will also note that with a serious heart condition or the building up of problems with the arteries, there is always a long term cure, a long term preventative, a long term correction of the condition. I would like to submit that if the schools are indeed in the condition of health that many reports are saying they are in, then it is going to require a long period of care and attention to put the schools into the health that we would aspire to during coming decades.

Because of this galvanic connecting of achievement test scores and the health of schools, we turn rather immediately to remedies which turn out not to address the health of schools. That is, they do not address the quality of educating in schools. And if the thermometers we use do not turn our attention to the quality of educating in schools, then the schools are not likely to get profoundly better, even if achievement test scores go up. And there is no question in my mind that achievement test scores in coming years will go up. They will go up particularly in the most mechanistic aspects of learning. And because of some of the reforms we are

beginning to think about, test scores will go up in some of the less mechanistic aspects of learning.

But I'm not at all sure that the quality of educating in schools will correspond to the rise in achievement test scores any more than the quality of education could be said to have paralleled the decline of achievement test scores -- about which we were concerned in the beginning of all of this.

I don't think it's entirely facetious for me to say that when the reports of 165 additional commissions are in, we already will have seen some of the signs of improvement. And I'm not at all sure that the implementation of the recommendations in those reports will make a very significant difference to the degree to which test scores are going to rise.

I made reference to 165 commissions -- that's the last report I've had. I've had to revise this number almost every time I've spoken on this issue. These are not casual bodies at work; they are state-level commissions. Most of their deliberations will lead to legislation which will be introducded in the sessions of the state legislatures this coming fall. However, we need to be aware that there are conditions having to do with the economy, having to do with the success of other institutions, and having to do with how we feel about ourselves that become immediately reflected in the schools. This does, indeed, cause us to turn to the schools in concern. We've not yet been very successful as social scientists in interpreting the reasons for the earlier decline in test scores. I doubt that we will be very successful in interpreting the increase in test scores in the years to come. It's part of the press around us.

Everywhere I go in the country, teachers are working harder. Students are working hard. Some students in high schools are thinking

about the law school they're going to attend or the post graduate work they are going to do after they complete their baccalaureate. There has been that kind of change. I'm not at all sure that it's more of an orientation of coming to grips with knowledge, but it's certainly an orientation of coming to grips with one's financial future.

As the test scores go up in the years to come, the rhetoric of self-congratulation on the part of those who are now making the recommendations will increase. That is, we will begin to adjust the rhetoric to the test scores and then say that what we're doing at the present time is improving our schools. And I'm raising some questions about such a connection with test scores.

Part of what is needed for a significant improvement to occur are comprehensive diagnoses of the educational enterprise and the educational condition. Yet in spite of all of the reports about schooling, there are still relatively few diagnoses. I want to present a perspective on these diagnoses and to deal with some specifics regarding their nature.

Let me turn first to the assessment of state responsibility. What should the state be doing at the present time? First of all, I think, states should articulate, much more clearly than they are currently doing, the comprehensiveness of our expectations for schools. And I don't think this is a capricious matter. We have data on these expectations. For example, in our Study of Schooling we looked at the expectations for schooling from a historical perspective. We analyzed the documents of all 50 states, we administered questionnaires to 8,600 parents, to 17,000-plus students, and to the teachers and principals in our sample. And what comes out of these data is that our society isn't backing off from a

comprehensive set of expectations for schools; society is still concerned about academic development, citizenship development, vocational development, and personal development.

Further, though James Coleman has been saying in some of his recent addresses that we can no longer think of the school in its surrogate parenting role, my conclusions are precisely the opposite. With demographics changing as profoundly as they are (in regard to the support of the home and the support of the religious institutions so far as they affect the young) we are expecting more of a surrogate parenting role of the schools than perhaps we did before. Those three institutions -- school, home, religion -- joined very closely when I was going to school. Now, more and more, we have deep concern about the school, in part because of the decline of the other institutions. It's interesting, for example, to note the number of parents in our sample who would opt for prayers in the school. And I'm not at all sure that this is only some kind of far-right religious concern representing a major turn in our society. I think it grows out of frustration on the part of parents (particularly with their young people entering puberty and adolescence) who, not knowing what to do about their children and hoping the school can do something, suddenly realize that teachers are human too. Therefore, it might not be a bad idea to have God in the classroom as well as the teacher, and so prayers become a pretty good idea. That may be an overstressed set of relationships, but I doubt that it's far afield.

It's interesting that only 37 of the 50 states articulate in any reasonably clear way the four areas of historic expectations -- academic, citizenship, vocational, and personal development -- that have emerged so clearly. It's interesting that California is one of the states that does

not articulate these expectations, but rather states goals in the context of the subject fields: education is teaching math, teaching science, teaching reading, teaching literature, rather than the using of those fields of knowledge for some higher human purpose (in addition to the purpose of learning the subject fields).

So, a state should be held accountable for the clear articulation of the expectations which careful surveys show are there. In addition, however, the state has a responsibility to define what the so-called common school means today. The common school was a vehicle in our society and part of its characteristics were designed to ready students for entry into the labor force. And until the early part of the century, the elementary school was the agreed-upon level of entry into the labor force and, as such, constituted the common school. Today one is expected, for entry into the labor force, to have matriculated from high school and have a high-school-leaving certificate. That means, then, that we should be evaluating the success of schooling not merely by the degree to which pushout programs (disguised as preparing the young for jobs, many of which are no longer there by the time they are to leave high school) increase your achievement test scores. Unfortunately, if you increase your achievement test scores while your retention rate is declining, your school gets brownie points.

But how about the criterion that the successful educational institution, K-12, is shaped like a rectangle? And that the most successful school is the one that keeps all those angles at 90 degrees? This means that those who begin in the kindergarten graduate with a high school certificate. However, the responsibility of schooling is not just to keep those young people in, but to assure comprehensive, democratic

access to the domains of knowledge that constitute a good general education. What a different criterion that would be. What a change that would bring about in regard to almost everything we do in schooling.

First of all, having a good school, as defined here, would require an enormous amount of collaboration among teachers and students. Students would have two responsibilities: one, to learn; the other, to help everyone else to learn. The best school would be the one that retains 100 percent of its young within a comprehensive program that we can agree to. And that means equity -- equity with respect to knowledge. But when we look at our data on tracking, it shows very clearly the disproportionate number of poor children in the low tracks and, in turn, the disproportionate number of minorities among those children. And when one looks further one notices the lack of equity in regard to the content in the upper and lower tracks. One also sees the lack of equity in regard to the pedagogical methods being used in the lower tracks versus the pedagogical methods being used in the higher tracks. And when one looks at teachers' expectations for the higher tracks which are clearly higher, clearly different, than teachers' expectations for the lower tracks, we find a monstrous situation of inequity, not the equity we wish to see.

The civil rights movement, once it resurfaces in this land with respect to education, will not be fought over access to schools. It will be fought over access to knowledge. And we will have to examine with great care those practices in schools which we take for granted, but which clearly operate against the principle of equitable access to knowledge for all within a comprehensive twelve or thirteen-year program leading to a high school certificate.

We all know the skillful ways in which we can subvert rewards for individual schools because of their gains in achievement test scores. For example, you can manipulate scores, either by leaving out groups of youngsters in the tests, or by the way you monitor those tests, or whatever. We need to pay attention to the work that Peterson is doing at McGill University right now, where he has begun to document the progression of youngsters through their aducational experience. He's going to spend twelve or thirteen years of his life at this -- documenting youngsters year after year longitudinally. And in talking with him just recently, he mentioned something he found just legion; the degree to which teachers provide subtle clues in walking around the room and watching the response of a youngster, to a tests. They say "Hmmm," and the child quickly looks again and changes the answer. We have all kinds of skillful techniques when the goal in mind is raising achievement test scores on the basis of those who are retained (particularly in these upper grade levels) rather than the extent to which children do well in a comprehensive curriculum and actually stay there until graduation.

Well, I spent more time on that than I intended to, but I want to give the notion now of how a different kind of quality indicator could be used by the state. I have <u>great</u> questions about the state's concern with individual children, and think the unit of selection for evaluation ought to be at a much higher level than that, such as the nature of the total program being offered. State responsibility should represent commitment to that broad set of expectations, commitment to the scope and breadth of the program and equity to which I just referred, and commitment to an evaluative framework commensurate with these expectations. And, in addition, the state must be commited to the development of quality

expectations in regard to the curriculum, its completion by all students, and the degree to which knowledge is humanized within that program for equal access of students. I'll come back to that point when I deal with the classroom or the school as the unit of analysis.

Let me turn now to the institution -- to institution-based or school-based assessment. Let's assume that our concern, at least in our rhetoric, is with the quality of educating in schools and with the health of schooling. Let's also assume that achievement test scores were never intended to measure the health of schools (some of you may have read, recently, the articles in the Los Angeles Times by David Savage and the quote from Gregory Amreg, President of Educational Testing Service, who says that the SATs were never intended to appraise the quality of educating or the quality of schools). Let me begin talking about institutions by referring to Sara Lightfoot's work.

Sara Lightfoot has published a book called Good High Schools. It consists of portraitures of six schools -- two private, two more or less upper socioeconomic class, and two urban high schools. She introduces a concept of "goodness". It's interesting to note that her concept of goodness deals so much with the degree to which the environment of the school supports the quality of living in those schools first, and the quality of educating in those schools after the quality of living has been raised to a point where instruction and learning can proceed. The schools are profoundly different and it's interesting that the Carver School in urban Atlanta is one of her "good schools." She talks about good "enough;" she doesn't say excellent, but good enough to be capturing the attention of far more students than was previously the case. And she describes a lot of things about that school that would make us wonder, on the basis of some

criteria, how in the world that could be a "good school" in her judgment. Then you begin to see Carver in some historical perspective, the lack of attention to the life of the school before the coming of a particular principal and a supportive superintendent, and the conditions in the school that operated against learning and the progress that had been made during recent years.

Schools have profoundly different cultures. There is no way to prescribe details in common for them. Indeed, in John F. Kennedy High School, another public school which Sara Lightfoot described in New York City, to prescribe in such a way as to seek to increase the intensity of academic life would simply be to increase those things in the culture which can be seen to be detrimental.

I urge you to read Lightfoot's book. It is a sensitive interpretation of life in six high schools. It is also interesting, for those of us who are interested in careful methodology, to read her commentary on educational research. She has some rather rough things to say about what we've been doing in the past, and admittedly she's defending work which she knows is going to be highly criticized in some quarters. Yet, it has not stopped her from moving in eight years from an assistant to a full professorship at Harvard, and she is now being sought after by several of the major institutions in the country.

We're beginning to get a different kind of handle on what is important in schooling and Lightfoot helps us a great deal. As a kind of a side comment, I'd like to note what Sara Lightfoot is talking about (after detailed descriptions of her six schools) or what Ted Sizer is talking about in Horace's Compromise (his analysis of teaching in schools and the

compromise that Horace had to make), is miles and legions away from where many of the commission reports are landing with respect to improvement.

Let me turn more specifically to what we might look at if we were concerned about the health, the condition of a school. My colleagues, Leigh Burstein and Kenneth Sirotnik, have been giving considerable attention to contextual analysis of schools, as have other colleagues at the Center for the Study of Evaluation at UCLA, and I think this kind of work is going to be very seminal. Leigh and Ken have done a lot of significant work and some preliminary publications are available; it is well worth considering.

What they're talking about is getting into the context of schools -- the conditions within schools. And when one looks at the conditions within schools, they take on meaning only as one relates that to a value system. And of course, that recognizes the fact that the understanding of schooling is in part a science, is in part an art. Because when it comes to the improvement of schooling, ultimately, we do that only through the application of norms, the application of values, the application of beliefs. But it also helps a great deal to take a look at present conditions.

For example, the degree to which a school has disruptive problems, the degree to which a school is torn apart by problems, can make it almost ridiculous to mount a staff development program based upon, say, specific cueing techniques for the improvement of instruction. When I visited one of the schools in our sample of high schools, getting the contextual sense to add to the hard data, I couldn't get to see the principal with whom I had an appointment at nine o'clock until eleven o'clock. He was on the telephone dealing with the courts all morning. In the mean time, I walked

through the school building with the vice principal. This person was Vice Principal for Curriculum and and Instruction and I asked him, "How do you spend your time?" He said, "Doing what I do now," as he reprimanded and separated students fighting in the hallway. He said, "My great frustration is that I came here because I was going to be Director of Curriculum of Instruction. Now I know why I came." And I looked at him and I knew why they had sent him. He was six-foot six, two hundred and thirty pounds, and an imposing figure as he walked through the hallway with another Vice Principal for Discipline who was about the same size. As they went through the hall, almost his entire time was spent in cleaning up fights. The major one that morning was within a group that had come in from the outside fighting with the students in the hallway. As we went around to the classes, they didn't bother to separate the children in the industrial arts classes, for example, to go into instruction with other children. They simply moved from working in the shop into algebra and mathematics and whatever, and the environment hardly changed. The conversation went on, the disruption went on, and one had to say that those children and those students were in no danger of learning anything that the school was trying to teach.

Sara Lightfoot points out as well that these are the problems that have to be addressed first. And so, in getting an assessment, in evaluating, if you will, the quality of life (what is the condition of the school environment) we discovered in our study of schooling the range of serious problems ran from none to a couple of problems, rated by teachers, parents, and students, as only mildly important or difficult, all the way to a school that had twenty-five problems which teachers, students, and

parents rated as very serious. Where do you begin improvement in that latter kind of school? Do you say "We're going to have a staff development program to improve instructional methods?" when the teachers aren't even conversing or communicating with rowdy, unruly students? Of course not. You begin where the culture of the school is. What are some practical problems you look at? How about time use, for starters, now that we're getting so much research on the importance of time in schools? We discovered in elementary schools with roughly the same length of school day, some children being in danger of learning what the school was trying to do for only eighteen and a half hours a week, and at another school children having 50 percent more instructional time, or twenty-seven and a half hours a week. I -- with some of my colleagues -- was one of the early people to speak with the National Commission on Excellence in their hearings. At their first morning of hearings, during the fifteen or twenty-minutes I had for testimony, I said, "I hope that one of the things you will not do is recommend increasing the length of the school day." Well, so much for expert testimony.

My reason for that was the climate of the school with eighteen and a half hours a week -- an obviously careless one with respect to the use of time (slow getting started, tardy children getting tardier while they waited to see the principal, recess stretching from fifteen minutes to thirty minutes, lunch hour dragging through much longer than was intended, and good old clean up time). There are enormous differences in the use of time, and these are clearly cultural problems in the school environment. These problems need to be addressed by the parents, the teachers, the students, under the leadership of the principal, in order to get enough

time to have a comprehensive curriculum. In analyzing our data, we concluded that in our sample of elementary schools, children were in instruction during the week for an average of twenty-two and a half hours. I looked at that time figure and laid it up against a model of access to the domains of knowledge in the elementary school and concluded that it was not enough. I didn't recommend a longer school day. I recommended that the local school work on that problem because they had enough time if they didn't spill so much of it. With twenty-five hours a week, for example, you've got ninety minutes a day of reading/language arts, an hour a day of math, fifty-five minutes a day of social studies, fifty-five minutes a day of science, three art periods a week, and health and physical education every day. With only eighteen and a half hours a week, you've got ninety minutes a day of reading/language arts, an hour a day of math, twenty-three minutes a day of social studies, thirteen of science, no art, and not much physical education or health. With twenty-seven and half hours, you've got the curriculum I just recommended and a lot more.

How about school climate? Do we not have climate indicators that we could use to determine, for example, what is valued most in the school culture? Friendships? Athletics? Smart students? Classes? Teachers? Or drugs? Alcohol? Games? Sports? Etc. As you know, the Select Committee on Education in Texas has been tackling this with a meat ax. They have concluded that there will be no athletics during the week -- this a recommendation coming from the legislature this June -- no athletics whatsoever on any weeknight, and no school-sponsored activities after six o'clock in the afternoon. They have prescribed a whole array of things because they're concerned that there is so much that is not close to the

learning process. Some of this bothers me a good deal, because I think it is possible, using time well, to have a comprehensive curriculum wherein students who are in vocational education programs may be getting the satisfaction and stimulation they need to perform in some of the other areas. Moreover, vocational education programs may be the entry into mathematics and science and the like for students who are getting turned o f. Notice that I used the word "education," however, and did not use the word "training." I'm talking about the kind of thing that John Dewey was doing with woodworking in his laboratory school at the University of Chicago.

I want to go on in this assessment. How do we get at the principal-teacher relationship? And then, from the research on effective schools and from elsewhere, what might be that most effective kind of relationship? It's very interesting that when we divided our schools up into the most satisfying quartile and the least satisfying quartile (using an index of satisfaction based on data from teachers, parents, and students), every single elementary school principal in the least satisfying quartile said the teachers are part of the problem. In the most satisfying quartile, I believe only one principal said that the teachers are part of the problem. I don't think that these were profoundly different people. And incidentally, when we looked at the correlations among satisfaction, school climate, class climate, principal-teacher relations, school-community relations, and the like, it was very clear that the most satisfying schools had a bond of trust and support and a working relationship between principals and teachers that was quite different from those in the bottom quartile.

Having assessed these things in the environment of the school, one still does <u>not</u> have a program of improvement. But now one can bring to

bear the value system of the professionals in the school, as well as interested citizens who are brought into what Bruce Joyce calls "the body of responsible parties". They can then begin to engage in long-term planning by saying: What is our first agenda item, second agenda item, third agenda item? And that becomes the agenda of improvement for the local school, approved by the superintendent and the board, and supported by them. This would result in such a different environment for school improvement than what is usually the case.

During the last fifteen months or so, I've had an opportunity to take another look at Edmonton, Alberta where, nine years ago, the superintendent of schools and the board introduced what I'm talking about -- a planning process with "every tub with its own bottom." Responsible parties at the level of the local school engaged in assessing their needs (in a primative fashion, admittedly, because we don't yet have the technology) and came up with priorities. They were able to sit down, in a non confrontational situation with the superintendent and the board, to review what it was that they were about and what they wanted to do. And they went about getting the endorsement of the superintendent and the board, getting differential support, getting funds for what they wanted to do, and then going about the business of doing it and reporting their progress the following year.

When I was back there a year ago, Edmonton had just been through a severe budget cut comparable to the budget cuts that have occurred in some districts around us. I expected to walk into a terrible morale situation -- teachers upset, principals upset -- a real downer. However, I walked into a very positive situation because here the superintendent and the board had called in all the principals and said, "We have to do a budget

cut of so much percent. Go back and revise your plans and see what you can do about cutting." All those principals came back several months later. They'd revised their plans; not only had they effected the budget cut, they now had a surplus. And then they asked the question, "May we keep it?" How foolish the superintendent and the board would have been had they not so permitted.

How different this is from a board of education obsessed with its importance, tearing its hair at one-thirty in the morning, reporting to workers how tired they are because they were fulfilling their responsibilities to the local community the night before, and slashing whole chunks out of the school program to nobody's satisfaction. In contrast, the smooth and morale-building process that occurred in Edmonton permitted, low and behold, good morale <u>while effecting a budget cut!</u> This, I think, is about the ultimate in concept. They were a long way, however, from being able to do this in a precise way, because we don't have the instruments, we don't have the technology, and we won't get them until we're concerned about such assessment.

Let me conclude with some brief comments on instructional assessment. Every bit as important, perhaps more important than whether or not a teacher produces attainment on an achievement test score, is the matter of whether or not a teacher in the classroom provides the students with an array of learning experience commensurate with our expectations for schooling. Do children ever engage in solving real problems? Do they ever have to work on a problem where there is no reward, until every member of the group has done his or her part, with or without the assistance of others, and the entire task is done by the group? After all, building that kind of

collaboration is the way we work in many aspects of life. In spite of the fact that our expectations for schooling talk about learning cooperative behavior, what we find in most schools is anything but that. We find from the beginning that learning in school has been learning alone in groups. To what degree do students do anything that requires some kind of response, some kind of product that's not preordained by the textbook or the workbook? To what degree do youngsters engage in modes of inquiry commensurate with what we think learning is? I'm not going to pursue this any further because one of the speakers will be doing that today, I'm sure, but I want to touch just briefly on the notion that there are more things to evaluating the effectiveness of the teacher than the product of achievement test scores.

What about class climate? Does class climate reflect what we know regarding human learning? We know a great deal, and clearly we won't reflect all of it. But is there some reflection there of what we know? One of the things we discovered in our studies is that there is very little variation used by teachers in the mechanics of teaching. The technology doesn't differ much from class to class to class. It gets to be terribly dull and boring as Kenneth Sirotnik has pointed out in his paper recently in the Harvard Education Review. But we did find that the climate surrounding this pedagogy differed quite markedly in the classroom. And, consequently, that there were classes that had more guidance, with the feedback that is one element of good teaching, as many researchers propose.

And then, finally, in this area, what about the assessment of the students' own experiences with school? What about those declining academic self-concepts where many students by the fourth grade are clearly saying,

"I'm not doing very well in school. I don't do well in mathematics and I don't feel very good about that." And then the need to recognize the change from focus on the school and focus on the subject that some of the tenth grade students in our sample indicated by saying, "Sometimes I don't feel good about myself at all." Is this the product we want of schooling? Isn't it interesting that we couldn't ferret out many differences in attitude towards school itself between those who were adjusting well and those who weren't? But we could identify the feeling of turning on oneself. What a marvellous job we've done of placing this institution in a high level of significance so that the individual says, "My failure is due to myself, and I don't feel good about myself at all."

What about students' academic self-concepts as they move through different schools? What about the curricula that students actually experience, not the curriculum that's offered? What about criterion- and domain-referenced measures that will tell us the growth that students have made in writing a paragraph from the time they're nine until the time they're twelve? Or what about our concern with the fact that students' art products seem so imaginative and creative at the age of five and six and seven, and then seem to get so stereotyped as they move on upward? What about taking a look at those developmental kinds of things?

And clearly, if we're going to get a handle on schools and their improvement, if we're going to have schools and educational systems that are healthy a decade from now, we're going to have to take a longitudinal view. We're going to need entry measures. Where's the school <u>now</u> in its health? Where's the state <u>now</u> in its articulation of goals? Where is the state <u>now</u> in the degree to which it is encouraging the development of

assessment instruments that get at <u>all</u> the goals of schooling and not just the mechanics? And then, what is the progress, whatever the criterion, that students have made over a period of time? Again, I refer to Sara Lightfoot (because she has made such a profound impression on me) and note the extent to which she assessed a school not in the light of now/cross-sectional/immediate measures but in the light of the history of that school: what was it doing now to become a better place for learning than it had been the year before?

Using Educational Evaluation for the Improvement of California Schools

Elliot Eisner

I would like to start out by clarifying what I think evaluation means in the context of education. I think the idea of educational evaluation often gets confounded with a host of other concepts that really obfuscate its meaning and confuse both professors of education and practitioners. We tend to mix up the notion of evaluation with the notion of measurement; we tend to confuse testing with measurement and evaluation. What I would like to do in this presentation is to sort out these concepts, because I don't regard them as being identical at all.

Measurement is a way of qualifying information according to some convention, some standard. It does not make a judgment about quantity. If I say, for example, that this room is larger than that hallway to which it is adjacent, I am making a descriptive claim that talks about quantity, and that descriptive claim is based upon my estimation, my appraisal, my judgment of space. But, in no way is it a measurement of the space that is out there and the space that is in here. For me to measure this room means that I have to employ some kind of device with conventional indices that represent the space that I occupy here, that this room represents, and that a hallway represents. Measurement is a way of quantifying information; it is a way of quantifying information according to some conventionally defined metric. A meter is a bar of metal kept in Paris and it defines what amount of length a meter is. It is arbitrary. We could define it in many other ways.

It is possible to measure things without evaluating them. I could

measure the length of this room, the width of this room, and the cubic space in this room without making a value judgment about whether this is good or bad, or indifferent, or appropriate or inappropriate. I could make a measurement of this room to determine how much carpeting I need in the room. This is a description of a state of affairs, it is not an evaluation. I can stand on a scale in the morning and I can measure my weight, and if I say, "Oh, Oh," then I am evaluating. But, if I simply want to know my weight, I am using that measurement in order to get information.

Evaluation has to do with making value judgments, value judgments about something that we care about. In education we care about educational processes and the consequences of those processes. Educational evaluation has to do with applying educational criteria to a state of affairs so that we can make some appraisal and assign some value to what we see occurring or to its results. So, when we evaluate we make judgments about the value of something on the basis of some criteria. The criteria that I employ to evaluate wine are not the criteria that I employ to evaluate classroom practice or its consequences. I use the criteria out of the wine industry o my experience as a wine connoisseur (of which I am not). When you make an educational judgment, educational value judgments, about the quality of your schools and the quality of your teachers and what they are doing, etc., you are applying educational criteria. And with respect to educational criteria, there are a wide array of differences as to what constitutes virtue in education. The criteria issue is itself a debatable, discussable issue, and it has been discussed for over 2,000 years (and I don't expect it to cease this afternoon).

Testing is not evaluation; it is simply one way of getting informa-

tion. It is very often a way of getting information that you could get in other ways if you waited for it. The use of testing is a way of constructing a situation, creating a device, typically, that elicits a response which can be measured. Further, we can engage in educational evaluation (and we certainly do engage in evaluation in the course of our lives) without using measurement and without using tests. For example, you folks are evaluating what I am saying to you. You are making judgments about its clarity, about its cogency, and about its relevance, and there is nobody in this room who is giving me a test! That is, I am engaged in a performance. I am providing information and you are making an appraisal of it. And if people start dozing off, I will get some feedback. If people start walking away, I will make some judgments about my performance and I'll start to do something else.

The first thing that you ought to recognize, if you do confound testing, measurement, and evaluation, is that these are three independent processes: We can evaluate without giving tests; and we can test without measuring; and we can measure wihout evaluating; and we can evaluate without measuring.

What about testing in evaluation? Whether a test or a measurement is an appropriate vehicle for securing information about which you can make value judgments educationally, is partly a technical problem. But there is no question in my mind that the use of tests (which often are confounded with educational evaluation and which people see as the only legitimate way to evaluate educational practice and its effects) does in fact have an affect on the educational priorities and the educational climate of schools.

Consider, for example, the headline in a relatively subdued,

relatively conservative newspaper: "Seniors' Scores Drop in Statewide Testing!". Let me read you three paragraphs. "California high school seniors dropped again this year on the average in a statewide assessment test, but educators on the Peninsula believe that their students improved on last year's scores. While the scores of the individual high school districts on the Califronia Assessment Program will not be released until May 11th, statewide results were reported this week to the State Board of Education. All seniors in California high schools were required to take the 30 minute tests. They scored 62.2% correct in the reading category, a decrease of 0.9% from the previous year; 62.6% in writing, a decline of 0.4%; 69.4% in spelling, a drop of 0.1%; and, 67.4% in mathematics, a decrease of 0.3%." Now, people tend to read headlines. Those headlines begin to set up expectations. And, interpretative information, particularly for test information, is not provided.

As another example, consider this array of North County Elementary Schools by district, showing grade 3 and 6 academic achievement test scores in a three year comparison. Teachers and parents look at these indices and they make juddments about the quality of education on the basis of the information that is very often rank ordered, out of context, without interpretative information. That kind of information gradually begins to affect what school teachers teach and what administrators are urged to pay attention to. And that kind of information has consequences for the kinds of reforms that are being implemented in schools - reforms that are by and large described on the basis of achievement tests that are often developed outside of the school context and which may or may not have much curricular validity.

We have been doing a study of high schools during the past two years

in the Bay area at Stanford, and there are some manifestations that we see when we look at classrooms. We are not looking at classrooms by going into them for a 45 minute visit with an observation schedule; we are trailing kids in schools, we are shadowing them for a two-week period. The research assistants in this project go to school with the youngster and they stay with that youngster for one full week, one week off and one week on. So, they shadow youngsters from Monday, 8:00 o'clock in the morning through the entire day. Very often they stay with them after school in order to get a sense of the quality of teaching, a sense of what's going on in classrooms and a sense of what kind of expectations are provided, etc. We do the same thing with teachers. Our research assistants go to school with teachers and they will spend a full week in their classrooms. I dare to say there is nobody in this room who is a school administrator who has spent one full week in a teacher's class. In one of the districts we are associated with, four teachers in a high school have been released by the superintendent to trail or shadow students in their own school. So, for the first time, after teaching in the school for 20 years, teachers are having access to their colleagues classrooms; and for the first time they are getting a vision of the nature of that environment, the common place that school is for the kids they are working with. And, this has proven to be an extremely illuminating experience because it allows us to get a fresh perspective.

One of things that we see is a good deal of curriculum fragmentation. When a multiple choice or short answer test is to be used, it influences the ways in which the students prepare and the kind of information that teachers give to the students and the ways in which teaching takes place. We are seeing teachers who torpedo their own lessons. They do a very nice

job of teaching in the course of the period; but near the end they remind students what is going to be on the test, giving them the implicit message that the rest of what they were paying attention to is not really important. That is of grave concern. If you have a vision of education that includes a great deal more than what tests assess (and it certainly is a vision that I have), then we need to recognize the influences testing has on instructional practice - for example: reducing the curriculum to small units of instruction; developing accounting procedures to record student assigments; maintaining records that objectify scores at end of the semester thereby depersonalizing education and "permitting" the teacher not to be responsible for making a personal judgment (or a professional judgment) on the work that a youngster has engaged in.

We see a great emphasis on the use of extrinsic rewards for the work that has been produced, that is, communication to youngsters that what really counts is getting a positive payoff on the basis of performance. We all want positive payoff; the question is what kinds of "payoffs"? Are we doing the sorts of things in schools, for example, that will enable youngsters to internalize what they are studying so that what they study in school become a part of their cognitive and affective repertoire, enabling them to use the ideas and the skills that they get in the context of classrooms and in situations that extend well beyond the classroom?

What I think is extremely important in terms of educational evaluation (and that has the potential to improve the quality of schooling is the examination of classrooms as they operate in the context of schooling. Consider curriculum as an intention, something that you organize as a body of content - a set of activities in an order, for example, in which they are to flow. If you think about the curriculum, in other words, as plans

for action, as body of material, then that body of material can be evaluated in its own right. One can pay very close attention to the educational substance of what is being intended in the classroom. You can look at a science curriculum, you can look at an art curriculum, you can look at a history curriculum, and you can make (if you have the ability to do this substantive judgments about the power of those ideas, about their importance within that discipline, and about whether these are the significant notions that kids ought to be exposed to. How many youngsters in your high school districts would be able to provide a decent explanation for the notion of random mutation and nature selection? Could they take that idea and apply it to the social world as well as the biological world? Do they see the relevance of this notion in terms of their understanding of biology? Is that a part of what they encounter in the courses that they take? It may very well be that they do. My point here is that the plans that we make for teaching, the curriculum that we design, the concepts, the generalization, the sorts of activities that are going to engage youngsters in at schools, can itself be an object of evaluation. And, if that program has insulated teacher from teacher, that has created conditions in which it is very difficult for the people who teach to learn about what they are up to as teachers. Most of you folks here have gotten out of teaching to become school administrators, perhaps because of the descretionary space that became avvailable to you as a school administrator but that you were denied when you were a teacher. A teacher goes to school at eight in the morning and she or he is with those youngsters until the end of the day.

We have created a structure which makes it very difficult for teachers to understand and to get feedback on how they do their business. Consider

the following thought experiment: If you were to conjure up a system that would increase the probability that there would be no growth in teaching over the course of a career, what features would you generate in your mind to increase that likelihood? What would you do? Well, one of the things that you might do is to create no incentives for being excellent in teaching. You might make sure that teachers got virtually no useful feedback about what they are doing. You might create infrequent, in-service education programs, removed from the shcool and taught by people who haven't crossed the threshold of the school themselves for a decade. Then you might think you will do your duty to inspire teachers in your district by inviting John Goodlad or Elloit Eisner or somebody like that, to give heartfelt speeches to jack them up in September so that they can carry themselves through June. In other words, I am suggesting to you a hypothesis. The hypothesis is that after teachers aquire the skills necessary to maintain the classroom and cope with the predictable crises that emerge in classroom, after two or three years in the classroom, growth in teaching is relatively flat. We have not provided the conditions in our schools to enable people to do better at their jobs. Yet we seem to pursue the idea that somehow we can humiliate practitioners into excellence by the publication of the performance of their students. This seems to me a wrong headed way to go about the improvement of California education.

So what is needed? We need to face up to the fact that we need to restructure opportunities in schools for teachers and administrators to learn what it is that they are doing in schools in their classrooms. I think we need to conceive of the evaluator's role as an educational role. That is, educational evaluation can inform teachers about what is subtle, but significant in classrooms. To accomplish this we first have to achieve

a set of conditions in schools that will de-isolate teachers from each other so that they have access to each other. Secondly, we need to establish a climate of trust in schools where people are willing to make themselves vulnerable to the observations of their colleagues. It means that we need to prepare school administrators and teachers in a way that will enable them to become connoisseurs of educational practice, because the presence of an individual in a classroom is no guarantee that they in fact will see what is important in that classroom. And the development of our ability to perceive the subtle but significant events that take place in school is a necessary condition to being able to provide feedback to the people who work in classrooms, so that their own activities as teacher can change. We need, I think, to develop a language of description that is not limited to quantitative information. I think there are wonderful uses of quantitative information for some sorts of things, but not for everything.

Think about the wide range of forms through which we represent the world. We represent the world discursively, we represent the world poetically, we represent the world figuratively, we represent the world quantitatively, we represent the world visually, and we represent the world kinesthetically. Our culture and our cognition are much wider than the vehicles that we use in schools to represent what we see. We first have to see, we have to perceive, we have to penetrate what is going on in classrooms. But we need the leeway and the space to inform people who operate in schools (and who formulate educational policy) as to where the problems are and where the achievements are.

Many of the things that we are seeing in the schools are extraordinary in terms of their achievement; we are seeing some marvelous teaching. We

are not, however, seeing as much of it as we would like. What strikes me in looking at schools (and in reading the case studies that our research assistants are producing) is the extent to which worse than mediocre teaching can continue year after year. I wonder frankly what the administrators are doing about this, and I wonder who is paying the price for this mediocrity, and I wonder why it is allowed to continue.. I have no conviction or belief that the publication of text scores will improve the situation, because what these lacking teachers need is much more subtle, it is much more supportive, and it is much more complex.

What is needed is a conception of inservice education that does not send teachers to service stations two times a year, but which builds in the concept of inservice education as an ongoing part of what it means to be a professional teachers. How can we construct schools so that the inservice part, the learning part, is part of what it means to be there? Can we create places for teachers so that you would be happy to say to your son or your daughter, "Yes, be a teacher, it's a fine thing to do, it will not thwart your growth, you can use every capacity that you have, the top is unlimited." Can we create places like that so that we don't have reservations about it?

I got out of it. I taught in a school, and I looked at my colleagues after two years of teaching in a high school of 3600 students, and, I said to myself, "I don't want to be in their place 25 years from now." So, I found a place where I had more space. And, so did most of you. We are not going to improve the educational lives of youngsters until we are able to provide more professional space for teachers. I don't think schools are going to be any better for kids than they are for the people who teach them. And the problem is to construct this kind of professional

environment. The problem is to design that structure and to communicate to people who have simple ideas about the improvement of education that those well intentioned plans may infact exacerbate the problem rather than ameliorate it.

Unfortunately, however, we are voiceless. Both professors are voiceless and school administrators are voiceless. Professors have a lessor right to be voiceless, but we are. We tend to be preoccupied with technical matters. And you are utterably vulnerable. When I talk about educational evaluation in schools I don't mean having a resident educational critic who goes around to classrooms and writes educational criticism. The model in my mind is to create school environments in which teachers can have access to each other and supportive and informative colleagues. How can you do that? What kind of substitute help can you provide to alleviate teachers of some of their responsibilities so they can have access to each other? What kind of climate of deliberation can you create so that people understand how it is that they are teaching? Look, I have been teaching since 1956. And I have been teaching at Stanford since 1965. You know, in the 19 years (or whatever it is) that I have been at Stanford, not ever has there been a colleague of mine that has come into my classroom to watch me teach! Not ever has a peer told me what I'm doing and what I'm not doing. I mentioned this to an audience once and one of the people in the audience said, "Well, Professor Eisner, why didn't you ask?" Why didn't I ask?...I didn't think of it. And the fact that I didn't think of it says a lot about my own professional socialization. It is not a part of what we do. You see, dancers have mirrors in the rooms in which they practice. Why do they have mirrors? Because they get information about how they move. Where are the mirrors in our classrooms?

The reflections in the students eyes are not good enough. And what we wind up with is trying to figure out (on the way home or on the way to work) how it went and why it didn't go as well as it did, or if it went well, why it went so well. And we never know whether what we think is what in fact took place. We haven't created a structure for it.

So, what am I saying to you? I am saying to you that I think we have grossly underestimated what it is going to take to improve California education. We cannot bully the schools into quality education. We need to give people a stake in what it is that they teach. The good school will expand individual differences rather than diminish them. And, we need to have programs which are diverse and which use multiple criteria even if it makes situations which are incommensurate. We ought not to allow a technology of testing provide ceilings on our aspirations and our intuitions and our insights. And, we need to create a climate of education, a structure of schooling in which the growth of the teacher is possible. Because it is through the teachers growth, through the teachers growing capacities, to appreciate what he or she is doing that the opportunities for educational experience are going to be defined for the young. Unless and until we face up to that task, we are going to be reeling from one mandate to another, making accomodations that deal with the superficial. An then years down the pike our successors will be doing the samething unless we face up seriously to what is needed in schools.

Teachers need to have a stake in their own operations and their own professional commitment. They need the time, they need the resources, they need twelve months' salary to plan, to deliberate; then they need an afternoon in which they can think with others about what's happening. They

need access to each other. We very badly need to find ways to convey to the public what it is we are achieving and what it is that we are not achieving, that we would like to achieve. I hope that a group like this could be the start of something that might be called "California Coalition for Quality Education" that would find the voice that I think is now absent in California education. I think we need in this state a group that can appropriate mandates for improvement of educational practice. I think we need to create a vehicle that in some way restore to our profession some modicum of authority and control within the districts for which we have responsibility. That's going to be hard when 80% or more of the funding is coming from someplace else. But, I think that is what is needed.

Some may view my ideas as impractical, but it strikes me that the greatest impracticality is to embrace procedures which in the long haul won't work even if they are "superficially practical." I would rather reach for something that I don't believe in, in order to accomodate the expectations of others. You have a very difficult task ahead of you. I can only wish you well in your effort and say to you that as far as I'm concerned, I am prepared to provide whatever assistance, whatever voice I can, in what we all know is perhaps one of the most important enterprises in the state. Thank you very much.

The Influence of Testing on Teaching and Learning

Norman Frederiksen

Speech given at a conference sponsored by the Laboratory in School and Community Education and the Center for the Study of Evaluation of the Graduate School of Education, University of California at Los Angeles, June 7, 1984.

In the first part of my talk, I'm going to argue that most current standardized achievement tests have serious limitations with regard to the skills and abilities they measure, and that these limitations may similarly limit what is taught and what is learned in school. I believe these effects are becoming more serious because of the growing use of standardized tests in school assessment, particularly the use of state-mandated minimum competency tests that are intended to set higher standards for promotion or graduation. I shall review some of the evidence to show specifically the what, how, and why of these effects.

In the second part, I plan to describe some testing methods that do not use the multiple-choice format and that might get at abilities that are not adequately assessed by most standardized tests. I shall also describe how tests that allow the student to write his/her own answers might be scored more accurately and economically than the usual essay test and how such measures might be used in schools to facilitate and improve the instructional process by encouraging the generalization of skills to new contexts and situations.

There is little question that tests do influence what is taught and

what is learned. The mere expectation that a test will be given tends to increase efforts to learn. Furthermore, the student's preparation for a test will be guided by his or her expectations as to what will be required by the test. That is the reason students often ask "What will the exams be like?" Students adopt different study methods for different test formats; if a multiple-choice test is expected, they will try to learn factual material, and if an essay test is expected, they will be more inclined to look for broader concepts and their relationships. Such differences in study methods are educationally important, and the net effect may be substantial, in view of the huge number of multiple-choice tests that students are required to take nowadays.

The number of multiple-choice tests given to school children each year has grown enormously over the past 25 years or so. Almost all the 50 states now have testing programs of one kind or another, and they typically involve multiple-choice tests. The number of published tests, such as the Iowa, California, and Stanford achievement tests, that are administered each year is estimated to be about 30 million. Furthermore, no one knows how many locally constructed multiple-choice tests are given as weekly quizzes and midterm and final exams each year.

The trend toward using tests to hold schools accountable has increased the influence of tests still more. In a school accountability feedback loop, information about a school is communicated to the school's constituencies--parents, potential employers, and even legislative bodies. Feedback to the school takes a variety of forms; parents can complain to the principal, employers can write letters to the editor, and the governing body can enact legislation. The loop is completed when the school administrators respond to the feedback by altering the curriculum,

retraining or reassigning teachers, or asking for more money. A good many state legislatures have in fact enacted laws mandating the use of minimum competency tests in order to set higher standards of achievement in school.

As a result of such pressures, scores on minimum competency tests have been on the rise. In my state of New Jersey, scores on the Minimum Basic Skills tests have increased slowly but steadily over the past few years. It is easy to see why, if you study the legislation. The program in New Jersey requires that rosters of test scores be released to all school districts, buildings, and classes, and that individual score reports be issued to students and their parents. General reports in the press are mandated. A list of the skills measured by each test is sent to teachers, and they are encouraged to use this information in their teaching. Old forms of the tests are made available for "appropriate instructional purposes"--which might turn out to be coaching. Schools failing to meet standards are subjected to review, and recommendations for remediation are prepared. If accountability feedback loops are not working in New Jersey, it is not the fault of the legislature.

Any improvements in the basic skills that result are, of course, much to be desired. My concern, however, is that the increased effort to teach the minimum competency skills decreases efforts to teach important abilities that tend not to be measured with multiple-choice tests.

A recent report of the National Assessment of Educational Progress (NAEP) suggests that there is indeed such an effect. NAEP's 1982 report showed that over the most recent decade performance on test items that measure the basic skills had not declined, but there had been a gradual decline in performance on items that measure more complex cognitive skills. For example, in mathematics it was found that 90% of 17-year-olds

could handle simple addition and subtraction; but for items that required problem solving, the decline was from 33% to 29%. Similar results were found for science, reading, and writing. In the case of writing, 75% of the 17-year-olds could write sentences with few mechanical errors, but for writing tasks that required analytic and logical skills, the percentage of writing samples judged to be "competent" dropped from 21% to 15% over the 10-year period.

Please understand that I am not trying to discourage the use of tests to influence instruction. On the contrary, I am all in favor of using tests to motivate and guide learners and their teachers, and even to provide practice. But we should be using tests that measure not only the basic skills but also the ability to process information rapidly and accurately, to apply principles in new situations, and to solve problems in forms they have not encountered before. Use of such tests, I believe, would help to improve instruction broadly, not just in the very basic skills that are easy to measure with multiple-choice tests.

Anyone who has prepared a multiple-choice test for a class must realize that it is indeed much easier to write items based on factual information involving names, dates, definitions, and formulas, than items requiring more complex cognitive operations. However, there have been few careful studies of the influence of test format on the behavior of the item writer. I can cite two.

One such study involved one of the Graduate Record Examination Board tests, the Advanced Psychology Test, which is a multiple-choice test given to college seniors who are applying for admission to graduate school. Members of a panel of 5 psychologists were asked to make a judgment about the kind of ability predominantly involved in responding to each item.

Definitions of four abilities were provided: memory, comprehension, analytic thinking, and evaluaion. Memory was defined as "simple reproduction of facts, formulas, or other items of remembered content." The consensus of the judges was that a large majority--70%--of the items were in the memory category, while 15% measured comprehension, 12% required analytic thinking, and only 3% involved evaluation. And this was a professionally made test that is widely used in admitting students to graduate schools.

Another study was based on a multiple-choice test intended to measure competence in orthopedic medicine. Judges were trained to sort the items into categories similar to those used in the GRE study. It was found that more than half of the items were unanimously judged to require only recall of information, while fewer than 25% were believed by even one judge to require interpretation of data, application, or understanding of a principle. An effort was then made to improve the next test by training the item writers to write items requiring the more complex cognitive processes. It was found that 50% of the items in the new test were still judged to require only recall of information.

These studies suggest that the difficulty of composing multiple-choice items that measure skills other than remembering may be a major reason for the tendency of multiple-choice tests to emphasize mastery of factual material.

Another line of research is concerned with how and to what extent taking a test influences student performance. Numerous studies have demonstrated that the expectation of a test increases test scores, and that taking a test tends to increase retention of the material tested. These effects are quite specific to what was tested; however, there is little

generalization. There is some evidence that free-response formats, such as short answer or completion tests, are somewhat more likely to improve retention. But such differences are not dramatic.

Other researchers have used the technique of inserting test-like questions in assigned readings. These studies confirm the finding that answering questions improves subsequent performance. But only factual items or questions were used in these studies.

Other studies of the effects of interpolated questions in text are more interesting because the effects of different kinds of questions were compared. One kind of question required verbatim recall of material in the text, and a second kind required more complex mental operations such as applying a principle in solving a problem. It was found that when the questions required the students to apply principles and to combine concepts and rules in solving problems subsequent performance improved substantially and generalized to new situations; and performance on verbatim questions did not decline.

A third line of research involves comparing tests presented in free-response and multiple-choice form with regard to the kinds of abilities they require. In such studies researchers have typically begun by choosing multiple-choice tests and then constructing parallel free-response tests by removing the multiple-choice options and replacing them with blanks in which students can write their answers. Then both types of test are given to samples of students, and various kinds of statistical analyses are made to find out if format makes a difference in what the tests measure. Several such studies have shown that format makes little difference.

Such research may be criticized on the grounds that the comparisons

involved only items that already existed in multiple-choice form. Parallel studies are needed where we begin with free-response tests intended to measure higher level cognitive abilities and construct parallel tests in multiple-choice form. Such comparisons have been made by several of us at ETS.

We began with a test we call Formulating Hypotheses. The problems were of a kind frequently faced by scientists. Each problem consisted of a brief description of a research study, a graph or table showing the results, and a statement of the major finding of the study. For example, in one problem a table showed that habitual users of marijuana improved in their visual-motor coordination after smoking a marijuana cigarette, while nonusers showed poorer performance. The task was to write hypotheses, or possible explanations, of the finding. Multiple-choice forms of such problems were constructed by providing a list of hypotheses from which the student could choose those he/she considered important. Scores were obtained that reflected the quality, number, and unusualness of the hypotheses that were written, or those that were chosen from a list.

It was found that correlations between corresponding scores for the two formats were very low. For example, for scores reflecting quality of the ideas the correlation between formats was .18, and for number of ideas, the correlation was .19. It appears that the two formats do not measure the same abilities.

In order to find out more specifically what abilities were involved by the two formats, the relationships of the scores to measures of several known abilities were investigated. These abilities included reasoning, verbal ability, knowledge of the area relevant to the problem, and ideational fluency, which may be interpreted as skill in searching for and

retrieving relevant information stored in memory. The most striking difference involved ideational fluency; none of the scores from multiple-choice versions was related to fluency, while for the free-response form scores reflecting number of ideas, number of unusual ideas, and number of ideas that are both unusual and of high quality were substantially related to fluency. Only the free-response form required a broad search of long-term memory for relevant ideas.

A similar study was carried out with a more elaborate problem-solving test that requires the student to go through a number of steps in seeking a solution to a problem, beginning with formulating hypotheses. Then the student is asked to indicate what information he or she needs in order to test the hypotheses or to suggest new ones. Then new information is provided, and the student revises his list of hypotheses. The student goes through half-a-dozen such cycles until he or she finally decides on a solution. Again, it was found that for problems posed in free-response form, the ability most involved is ideational fluency, with reasoning involved particularly at steps where inference is required; for the multiple-choice format these relationships were all substantially lower. Thus, it again appears that the multiple-choice format does not require the same skills as the free-response format.

The research I have briefly reviewed I think supports three conclusions about how test format influences behavior:

First, test format influences the kinds of items that a test maker writes. Because it is much easier to write multiple-choice items that measure factual knowledge, the item writer tends <u>not</u> to write items that measure skills in analysis, problem solving, application of principles, and the like--even when they try hard.

Second, tests do influence student performance. If the free-response tests are adaptations of multiple-choice tests, format makes only a small difference. But evidence from studies of the influence of questions interpolated in text indicates that questions that require complex cognitive processing, in contrast with factual questions, do improve performance on subsequent tests, and there is transfer to other kinds of problem-solving tasks. Similar results might be expected for items incorporated in tests.

Third, research on the influence of format on what abilities the test measures indicates that format makes little difference if one compares multiple-choice tests with their free-response counterparts. But if one begins with free-response tests that require complex cognitive processing and compares them with similar tests cast in multiple-choice form, format strongly influences what is measured. In particular, ideational fluency is important only if the student has to compose his/her own answers rather than choose them from a list.

Now let me turn to the second part of my talk by considering some alternatives to multiple-choice tests. I shall comment first on essay tests, and variants of essay tests that can be scored more objectively. Then I shall consider a variety of testing procedures that have little resemblance to conventional tests.

We are so accustomed to multiple-choice, true-false, and completion tests that we seldom consider other possible formats. The usual alternative is an essay test. But teachers don't like essay tests because grading is onerous and time consuming, and test publishers don't like them because they can't be scored with a machine. Another problem is low reliability of grading. In one study, 300 essays written by college

freshmen were graded independently by 53 experts, including English teachers, editors, writers, lawyers, and scientists, using a 9-point scale. It was found that no essay was given fewer than 5 of the 9 possible ratings, and 34% of the essays were given all 9 of the ratings. Essay grades may depend more on who reads the essay than on who wrote it.

One way of achieving higher reliability is to use several readers instead of one and to pool their judgments. Since this is a pretty expensive way to grade essays, a method called "wholistic" scoring has come to be used. In this procedure the essay is graded quickly and impressionistically by two or more readers. This brings down the cost, but it is certainly not possible to state very precisely what the grade means.

No method of scoring that involves people rather than machines can compete with the multiple-choice test. But there are methods of evaluating written protocols that may turn out to be faster, less expensive, and more reliable than the usual method of grading essays, and the method can provide not one but a number of scores that have very precise meanings. Such methods would not work very well for such essay topics as "How I spent my summer vacation," but they would probably work for assignments that are well structured in the sense that all the students are attempting to accomplish the same task by more or less similar procedures.

The Formulating Hypotheses test that I described earlier is an example. I mentioned the names of some of the scores, but I did not describe the scoring procedure.

We call the method <u>category scoring</u>. Several preliminary steps are required. The first step is to develop a classification of the ideas produced by a sample of students in response to the problem. In the case of Formulating Hypotheses, these ideas are the hypotheses that the students

thought might account for a research finding. Our procedure for classifying responses is to copy each hypothesis on a 3x5 card, then sort the cards into piles that contained identical or closely similar ideas. Initial agreement among sorters was generally quite good, and after discussion a consensus was reached on the number and nature of the categories. Then a definition was written for each category, trying to differentiate clearly each category from all the others.

The next task is to ask a panel of judges to make an evaluation of the quality of each response category. Then a quality value is assigned to each category on the basis of the combined judgments.

The scorer's task is then comparatively simple--to read each hypothesis and match it to one of the categories. Scorers do not have to be experts to do this. After a reasonable amount of training and practice, agreement between scorers is good. The category assignments are entered into the computer, along with the quality values and information about the frequency of occurence of each category. A variety of scores can then be generated. We used scores that reflected the number of ideas written, the number of good ideas, the average quality of ideas, the number of unusual ideas, and the number of ideas that were "creative" in the sense that they were both unusual and of high quality.

It is also possible to ask the panel to make other judgments about the ideas as a basis for additional scores. For example, a hypothesis might have been directly suggested by information in the problem statement, it might have resulted from inference based on such information or, if it was unrelated to any information given, it must have come from a search of long-term memory. Scores to represent the number of ideas from each source can easily be generated by the computer.

There are many possible applications of category scoring. We have used it to score medical problem-solving tests, which are paper-pencil simulations of a doctor's encounter with a patient, as well as for other tests of scientific thinking called Solving Methodological Problems and Evaluating Proposals. Although the method works best when the problem constrains all the students to respond in ways that are roughly similar, the method might be applied even to essays, if the topic assigned is very clearly specified. A content analysis of a sample of essays on a given topic might reveal a common core of ideas, and relationships among ideas, that could be sorted into categories, which in turn could be evaluated and used as the basis for a scoring system.

The test could be used for instructional purposes by having each student score his or her own protocol. When the problem is completed, the student could be given the category definitions and told to match his or her responses to the categories. Then feedback could be given in the form of the quality values, along with a critical statement of the good and poor features of each category. If a large enough number of such test problems is available, a substantial amount of practice could be given, and if the problem settings are realistic and varied, such practice should promote generalization and encourage learning by discovery.

So much for scoring the protocols of free-response tests. Now let us consider some ideas for testing that grow out of theories of cognition. These ideas are quite different from conventional tests. You may even think some of them are wild ideas.

One such idea has to do with measuring speed in performing cognitive tasks. The trend over the past 20 or 25 years has been toward <u>power</u> tests as opposed to <u>speeded</u> tests. The most important reason is probably the

desire to be fair. The student with a low score who could have gotten all the items right if he/she had had more time may feel that he/she was cheated. Actually, we do not have to choose between speeded and power tests--we can give both. Let me explain why I think it is important to measure speed as well as power.

The reason that speed is important has to do with certain attributes of memory. Cognitive psychologists distinguish several kinds of memory, but I will discuss only two, called long-term memory and short-term or working memory.

Long-term memory is the limitless and relatively permanent repository of one's knowledge. It contains a huge amount of information, including knowledge of procedures as well as facts and their relationships. We are not aware of any of this information, however, until some part of it is transferred to working memory. Working memory contains the information we are aware of and are actively using at a given time. The term information processing refers to the flow of information into and out of working memory, by such processes as retrieving information from long-term memory; receiving sensory inputs; comparing, combining, and transforming items of information; and placing new or altered information back in long-term memory.

An important feature of working memory is that it has very limited capacity; it can accomodate only six or seven items of information at one time. Any information above this limit crowds something out, as you know if you have ever taken a digit-span test. This small capacity imposes a serious limitation on one's ability to deal with complex problems. But since we are able to deal with complex problems, there must be ways to compensate for the limitation. This is where speed comes in.

One method of compensating is called automatic processing. With a

great deal of practice, it is possible to carry out mental activities automatically, without paying attention and without using up the limited capacity of working memory. An example is one's ability to drive a car along a familiar route while carrying on a conversation with a companion. An example from the school room is the ability of a skilled reader to decode the symbols that comprise a word automatically, without paying attention, and thus without interfering with his or her ability to deal with more complex aspects of reading. Similarly, the mathematician can carry out elementary algebraic operations automatically, without attention, while attending to his more remote goals in solving the problem.

How can we measure the development of automatic processing skills? Cognitive psychologists assess automaticity by measuring latencies, or reaction times, in responding to simple tasks that are components of more complex skills. For example, a microcomputer might be used to present a list of words one at a time to a student, and to measure the latencies as he/she responds by saying each word as quickly as he/she can. Individual differences in latencies on such tasks may be substantial, even among students who make almost no errors in saying the words, and they discriminate between good and poor readers.

A simpler method of measurement that might be just as good, from an instructional point of view, is a paper-pencil test containing a long string of orthographic symbols, some of which are words and others nonwords. The task might be to mark as rapidly as possible the symbols that are words. The last item attempted before time is called would be the score. Similarly, tests might also be used to measure speed in carrying out other component tasks, such as filling in blanks to indicate the antecedents of pronouns used in sentences.

Automatic processing, then, is one way to compensate for the limited capacity of working memory. Another method, which is closely related, has to do with pattern recognition, which is the ability to perceive a pattern of related parts quickly and accurately. Like automaticity, pattern-recognition skills are acquired only through a great deal of practice. A chess grandmaster can look for a few seconds at a chessboard with the pieces in a midgame position and then reproduce on another board the positions of the 25 or 30 pieces almost without error. Ordinary players given the same task can place correctly only 5 or 6 pieces--a number which is more consistent with what we know about the capacity of working memory. What the grandmaster perceives is 5 or 6 chunks or clusters each of which is a pattern of 5 or 6 related pieces.

Similar results are found in other areas of expertise. Electronics experts can quickly identify the patterns in a circuit diagram that represent the elements corresponding to the power supply or a stage of amplification, and experienced physicians can recognize in a case workup the pattern of signs and symptoms that correspond to a diagnostic category.

How can we measure pattern-recognition skills in schools? As in the case of reading, both speed and power tests are desirable. Power tests are especially important at early stages in acquiring a skill, to find out about the number and kind of patterns that can be recognized. Speeded tests are important at later stages when through practice recognition is becoming automatic. Methods analogous to those used in assessing performance on the components of reading could be used in other areas, such as recognizing geographical features from contour maps, identifying organic compounds from representations of carbon chains, or locating body lesions from X-ray photographs.

Another testing idea suggested by theories of cognition has to do with how one represents a problem internally. Such a representation may take many forms. A word problem in mathematics, for example, may be seen by various students as a set of verbal statements, a chart or diagram, an equation or set of equations, or a procedural flow chart of some sort. An inadequate representation may make the solution of a problem difficult or impossible. How can we find out how a particular individual represents a problem?

This is a difficult question to answer because problem solvers usually don't know how they represent a problem; therefore, it must be inferred. A research method that has been used by cognitive psychologists is to present to students with fairly large set of problems from some domain, such as physics, and to ask them to sort the problems into sets that are similar with respect to how they are solved. Striking differences between students and experts are found. Students tend to sort the problems on the basis of surface features, such as pully arrangements or weights on inclined planes, while experts sort on the basis of the physical principles that are involved, such as Newton's third law. Tests based on such a procedure might reveal something about a student's stage of development in forming useful representations of problems.

Another important factor in problem solving is how information is stored in long-term memory. This is important because good organization of the stored information facilitates retrieval and enhances the likelihood of seeing interrelationships among the stored items of information. Making a test to determine how information is stored would appear to be impossible; but a beginning has been made. The method is to find out how key concepts in an area are interrelated by a given individual. In the area of

mechanics, for example, there are a dozen or so important concepts, including mass, density, velocity, acceleration, force, and so on. One can present to the student all the possible pairs of these terms and ask him to make a judgment about the strength of the relationship between the members of each pair. A statistical method, such as multidimensional scaling, can then be used to produce a cognitive map showing the dimensions of the system and their interrelationships. Such a picture could be compared with the analogous structure based on the judgments of experts. The cognitive map presumably reflects the student's understanding of a large interacting system of concepts at a certain phase in his learning, and it could be compared with similar representations obtained at earlier and later stages.

By way of summary: I have described several possible testing methods that with further development might be used to replace or supplement multiple-choice tests. Two of the ideas are concerned with skills that help one to compensate for the limited capacity of working memory; they are the automatic-processing and pattern-recognition skills. I suggested that it would be relatively easy to measure automatic processing skill in a particular areas of expertise, such as reading, by using speeded tests with relatively easy items. I believe it would be quite feasible, also, to measure skill in pattern recognition by similar methods, although we may need more investigation to identify the patterns that are salient for a particular area of instruction. I consider this kind of testing to be very important, because these are the skills that make it possible to attend to the more complex aspects of a problem or a situation without getting bogged down in the detail.

Methods for measuring how problems are represented internally and how

information is organized in long-term memory are also potentially important, but at the moment such measurement methods may be more important for the researcher than for the educator.

I also described something I called category scoring, which may make it feasible to use tests that elicit fairly lengthy written responses. It has been demonstrated that the method works quite well for devices like the Formulating Hypothesis test, but we need to find out to what extent it can be adapted to other formats. I consider this to be a very important development if it encourages teachers to assign more tasks that require constructed responses.

Another way of appraising the new test ideas has to do with the coachability of the tests. Some are coachable in the bad sense that coaching may improve the test score without improving the ability measured by the test. For example, students could be taught a "correct" cognitive map without altering the actual knowledge structure. But other tests may be coachable in the good sense that coaching for the test would also improve the ability measured by the test. I consider this a good feature of a test because the test can then be used as an instructional tool to provide the practice and feedback that are so necessary for learning.

It has been argued by Walter Doyle that tasks are the basic treatment units of a school, and greater emphasis should be given to task assignments such as writing papers, solving homework assignments, and taking tests. If the tasks are properly designed, they could help students to acquire not only the knowledge base but also the information-processing skills that are necessary for developing high levels of proficiency in thinking.

I suggest that the primary purpose of tests, tasks, scorable exercises, or whatever you want to call them, should be to provide practice

with feedback to students and diagnostic information for teachers. Taking such tests or exercises should be daily occurrences rather than something that happens once or twice a term for the purpose of assigning grades. Properly designed materials would help students not only to acquire competency in basic skills, but also to acquire high levels of proficiency in pattern recognition, automatic processing, and other information-processing skills that make it possible for students to advance to higher levels of accomplishment. And if the tasks assigned involve a wide variety of realistic contexts and situations, proficiency may generalize to the difficult real-life problems that will arise in the future.

All this may strike you as fine; but who is going to pay for it? It is certainly true that the tests I described cannot be scored at the rate of 10,000 answer sheets an hour. But I have a few suggestions that might help in terms of costs. One is that some of the tasks could be programmed for microcomputers, so that the computer could give the test, score it, and even provide comments and suggestions to the student. Another idea is that students might score their own tests, for prompt feedback. Another is that the material that is most costly to prepare could be provided by consortiums of school people and testing organizations for use on a wide scale. Finally, if we consider the administration and scoring of tests as <u>instruction</u> rather than assessment, the cost may not seem exhorbitant. And the usual testing for grading and assessment purposes can be dropped because better information will be available as a by-product of instruction.